PATENT ABSTRACTS OF JAPAN

(11)Publication number : 11-015832
(43)Date of publication of application : 22.01.1999

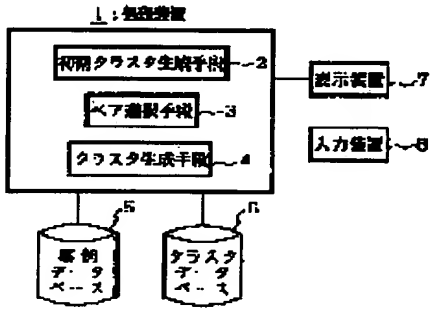
(51)Int.Cl. G06F 17/30

(21)Application number : 09-168267
(22)Date of filing : 25.06.1997

(71)Applicant : FUJITSU LTD
(72)Inventor : OOTA TADAKO
YUGAMI NOBUHIRO
OKAMOTO AOSHI
SATO OSAMU

(54) CLUSTER GENERATION DEVICE AND RECORDING MEDIUM

(57)Abstract:
PROBLEM TO BE SOLVED: To generate a cluster hierarchy permitting that one example belongs to plural clusters by means of shortening cluster generation time by considering the cluster whose number of appearing examples in common is large in a cluster set to be highly associated and generating the new cluster from the plural clusters.
SOLUTION: An initial cluster generation means 2 generates the clusters for respective attribute values based on the given example, which is read from an example data base 5. A pair selection means 3 selects the pair of the clusters in the generated clusters. A cluster generation means 4 counts the number of the examples contained in common on the selected pair, generates the new cluster from the pertinent pair of plural clusters in the largest case and stores it in a cluster data base 6. The cluster whose number of the appearing examples in common is large in the cluster set is considered to be highly associated. The generation of the new cluster from the plural clusters is repeated and the cluster hierarchy is generated.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平11－15832

(43)公開日　平成11年(1999)１月22日

(51)Int.Cl.[6]	識別記号	ＦＩ	
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ　15/401	３１０Ｄ

審査請求　未請求　請求項の数４　ＯＬ　（全９頁）

(21)出願番号　特願平9－168267

(22)出願日　平成９年(1997)６月25日

(71)出願人　000005223
富士通株式会社
神奈川県川崎市中原区上小田中４丁目１番１号

(72)発明者　太田　唯子
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内

(72)発明者　湯上　伸弘
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内

(74)代理人　弁理士　岡田　守弘

最終頁に続く

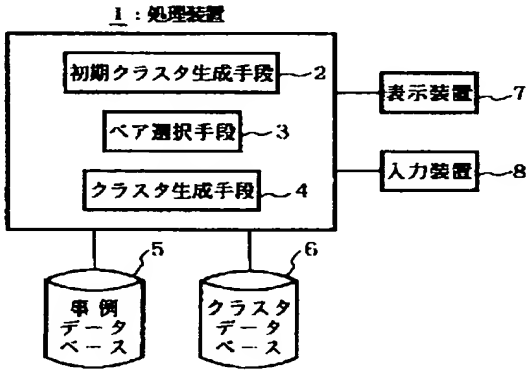
(54)【発明の名称】　クラスタ生成装置および記録媒体

(57)【要約】

【課題】　本発明は、与えられた事例をもとにクラスタを生成するクラスタ生成装置に関し、各属性の各値についてその値を持つ事例を集めてクラスタを生成し、クラスタ集合中で共通して出現する事例の数の多いクラスタは関連が高いと見做して複数クラスタから新たなクラスタを生成することを繰り返してクラスタ階層を生成し、クラスタ生成時間を短くして１つの事例が複数のクラスタに属することを許すクラスタ階層を生成することを目的とする。

【解決手段】　与えられた事例をもとに属性値毎にクラスタを生成する手段と、生成されたクラスタ中の複数のクラスタに共通して含まれる事例の数が多い場合に当該複数のクラスタから新たなクラスタを生成する手段とを備えるように構成する。

本発明のシステム構成図



1

【特許請求の範囲】

【請求項１】与えられた事例をもとにクラスタを生成するクラスタ生成装置において、

与えられた事例をもとに属性値毎にクラスタを生成する手段と、

上記生成されたクラスタ中の複数のクラスタに共通して含まれる事例の数が多い場合に当該複数のクラスタから新たなクラスタを生成する手段とを備えたことを特徴とするクラスタ生成装置。

【請求項２】上記複数のクラスタから新たなクラスタを生成するとして、複数のクラスタの和集合を新たなクラスタとすることを特徴とする請求項１記載のクラスタ生成装置。

【請求項３】上記複数のクラスタから新たなクラスタを生成するとして、複数のクラスタの積集合を新たなクラスタとすることを特徴とする請求項１記載のクラスタ生成装置。

【請求項４】与えられた事例をもとに属性値毎にクラスタを生成する手段と、上記生成されたクラスタ中の複数のクラスタに共通して含まれる事例の数が多い場合に当該複数のクラスタから新たなクラスタを生成する手段として機能するプログラムを格納した記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、与えられた事例をもとにクラスタを生成するクラスタ生成装置に関するものである。

【０００２】事例を分類するクラスが与えられていない場合、どのようなクラスに分類するかが重要な問題となる。クラスタリングは、そのような場合、クラスそのもの、または部分集合となるようなクラスを生成する技術である。一般に、クラスタは似た事例の集合である。例えばアンケートデータなどから、似た人達のグループを検出するのに利用できる。このクラスタの生成においては、クラスタ間の類似性の計り方、クラスタの生成の仕方によって、得られるクラスタ階層の性能が変わってくる。本発明は、そのクラスタ間の類似性の計り方、生成の仕方に関するものである。

【０００３】

【従来の技術】従来、クラスタ生成の手法には、逐次的手法と一括的手法とがある。逐次的手法は、クラスタリングされる事例を、一度ではなく、少しずつ時間をおって与えられる。新しい事例が与えられた時、それまでに作られたクラスタを始めから作り直すのではなく、それに変更を加える操作を行う。そのため、クラスタ生成に必要とされる時間は短いが、生成されたクラスタ階層の性能は、事例の入力順序に大きく依存する。

【０００４】一括的手法は、事例が一度に与えられるため、入力順序の問題はないが、クラスタ生成に時間がかかる。例えば、代表的な手法としてＷａｒｄ法があるが、この計算量は、全事例数の３乗である。

2

【０００５】

【発明が解決しようとする課題】上述したようにクラスタ生成を従来の逐次的手法で行うと、新しい事例が与えられたときにクラスタ生成を迅速に行うことができるが、生成されるクラスタ階層が事例の入力順序に大きく依存してしまう問題がある。

【０００６】また、従来の一括的手法で行うと、入力順序の問題はないが、クラスタ生成に時間がかかるという問題がある。また、逐次的／一括的手法に共通して多くの手法は、１つの事例が１つのクラスにしか属さないようなクラスタ階層が生成される。ある問題を解決するために、複数の視点が要求される場合も多く、従来の１つの事例が１つのクラスにしか含まれないような階層を生成する方法では、各視点ごとに階層を生成する必要も生じてしまう。そこで、事例の入力順序によらず、階層生成に必要な時間が短く、更に、１つの事例が複数のクラスタに属することを許すようなクラスタ階層生成手法が望まれる。

【０００７】本発明は、これらの問題を解決するため、各属性の各値についてその値を持つ事例を集めてクラスタを生成し、クラスタ集合中で共通して出現する事例の数の多いクラスタは関連が高いと見做して複数クラスタから新たなクラスタを生成することを繰り返してクラスタ階層を生成し、クラスタ生成時間を短くして１つの事例が複数のクラスタに属することを許すクラスタ階層を生成することを目的としている。

【０００８】

【課題を解決するための手段】図１を参照して課題を解決するための手段を説明する。図１において、処理装置１は、図示外の記録媒体からプログラムを主記憶にローディングして起動し各種処理を行うものであって、ここでは、初期クラスタ生成手段２、ペア選択手段３、およびクラスタ生成手段４などから構成されるものである。

【０００９】初期クラスタ生成手段２は、クラスタを生成するものである。ペア選択手段３は、クラスタのペアを選択するものである。クラスタ生成手段４は、選択したペアについて共通出現事例数を計数し最も多い数のペアを新たなクラスタとして生成などするものである。

【００１０】事例データベース５は、事例を格納したものである。クラスタデータベース６は、生成したクラスタを格納するものである。表示装置７は、クラスタを表示などするものである。

【００１１】入力装置８は、各種指示やデータを入力するものである。次に、動作を説明する。初期クラスタ生成手段２が事例データベース５から読み出した与えられた事例をもとに属性値毎にクラスタを生成し、ペア選択手段３が生成されたクラスタ中の複数（例えば２つ）のクラスタのペアを選択し、クラスタ生成手段４が選択されたペアについて共通して含まれる事例の数を計数し最

も多い場合に当該ペアの複数のクラスタから新たなクラスタを生成し、クラスタデータベース６に格納するようにしている。

【００１２】この際、複数のクラスタの和集合を新たなクラスタを生成するようにしている。また、複数のクラスタの積集合を新たなクラスタを生成するようにしている。

【００１３】従って、各属性の各値についてその値を持つ事例を集めてクラスタを生成し、クラスタ集合中で共通して出現する事例の数の多いクラスタは関連が高いと見做して複数クラスタから新たなクラスタを生成することを繰り返してクラスタ階層を生成することにより、クラスタ生成時間を短くして１つの事例が複数のクラスタに属することを許すクラスタ階層を生成することが可能となる。

【００１４】

【発明の実施の形態】次に、図２から図５を用いて本発明の実施の形態および動作を順次詳細に説明する。

【００１５】図２は、本発明の動作説明フローチャートを示す。これは、図１の構成の詳細な動作を説明するフローチャートである。図２において、Ｓ１は、事例集合を入力する。

【００１６】Ｓ２は、パラメタＫとクラスタ制限数Ｌを決定する。これは、右側に記載したように、パラメタＫとクラスタ制御数Ｌを入力し、これをもとに右側に記載したように、Ｋ×標準偏差＝標準偏差の正の実数倍の決定、およびクラスあるいはその部分集合（クラスタ）の総計の最大数をＬとして決定する。

【００１７】Ｓ３は、初期クラスタ階層を生成する。これは、Ｓ１で入力された事例集合について、後述する図３に示すように初期クラスタ階層を生成、即ち、後述する図３に示すように、入力された全事例を含むクラスタをＣrootとし、葉クラスタはそれぞれ属性ａｍの値がｖｍｖである事例を集めたクラスタとからなる初期クラスタ階層を生成する。

【００１８】Ｓ４は、終了か判別する。これは、Ｓ５以下の処理について、クラスタ数がＬ個となるか、ペアとして選択するクラスタがなくなるまで繰り返すことが終了したか判別する。ＹＥＳの場合には、終了する。ＮＯの場合には、Ｓ５に進む。

【００１９】Ｓ５は、クラスタペアを選択する。Ｓ６は、共通出現事例数を数える。これは、選択したペアのクラスタについて、共通して出現する事例の数を数える。

【００２０】Ｓ７は、推定共通事例数を計算する（後述する）。Ｓ８は、Ｓ６で数えた共通出現事例数がＳ７で計算した推定共通事例数よりも十分に大きいか判別する。ＹＥＳの場合には、Ｓ９に進む。ＮＯの場合には、Ｓ４に戻り繰り返す。

【００２１】Ｓ９は、事例数の割合を計算する。Ｓ１０は、一番大きい割合を記憶する。Ｓ１１は、全ペア試行したか判別する。ＹＥＳの場合には、Ｓ１２で一番割合の大きいペアを選択し、Ｓ１３で新たなクラスタを生成し、Ｓ４に戻り繰り返す。一方、Ｓ１１のＮＯの場合には、Ｓ４に戻り繰り返す。

【００２２】ここで、Ｓ３ないしＳ１３について以下に詳細に説明する。後述する図３に示すように、全ての事例を含むクラスタをＣrootとして生成し、葉クラスタとして属性ａｍの値がｖｍｖである事例の集合を生成する（Ｓ３）。そして、共通して出現する事例数のクラスタのペアの関連度を評価する。その数が推定共通事例数よりも十分に大きい場合（Ｓ８のＹＥＳの場合）、関連が高いと見なす。推定共通事例数は、以下の式によって計算する。

【００２３】

$$e_{ij}=(|C_i|\times|C_j|)/N \quad (1)$$

ここで、$|C_i|$、$|C_j|$はクラスタｉ、ｊに含まれるそれぞれの事例数を表し、Ｎは全事例数を表す。このときの標準偏差は、

$$d_{ij}=\{|C_i|\times|C_j|\times(N-|C_i|)\times(N-|C_j|)\}^{1/2}/\{N^2\times(N-1)\}^{1/2} \quad (2)$$

となる。ここで、パラメタＫを含む、以下のような式と見做され、選択候補となる（Ｓ８）。

【００２４】

$$|C_i\cap C_j|>e_{ij}+K\cdot d_{ij} \quad (3)$$

（３）を満たす時に、その２つのクラスタの関連が高い

ここで、左側は後述する図４の（ｅ）の積集合を表す。次に、下記の式（４）を計算する（Ｓ９）。

【００２５】

$$|C_i\cap C_j|/(|C_i|\times|C_j|) \quad (4)$$

ここで、左側の式は後述する図４の（ｅ）の積集合を表し、右側の式は後述する図４の（ｄ）の和集合を表す。

【００２６】これら計算した結果から、式（３）を満たし、式（４）を最大化するクラスタペアを選択することにより、事例数が多いクラスタ同士が選択される傾向を緩和できる（Ｓ９、Ｓ１０）。

【００２７】そして、選択されたクラスタペアの和集合を新たなクラスタとして生成し、選択ペアの上位にリンクすることにより、後述する図４の（ｂ）となる。また、選択されたクラスタペアの積集合を新たなクラスタとして生成し、選択ペアの下位にリンクすることにより、後述する図４の（ｃ）となる。

5

【００２８】以上のように、各属性値についてクラスタを生成し、関連の大きいクラスタペアから新たなクラスタを生成することを繰り返しクラスタ階層を生成することにより、クラスタ生成時間を短くして１つの事例が複数のクラスに属することを許すクラスタ階層を生成することが可能となる。以下順次詳細に説明する。

【００２９】図３は、本発明の説明図（その１）を示す。これは、初期クラスタの概念を説明する図である。図３の（ａ）は、初期クラスタの概念図を示す。ここで、全ての事例を含むクラスタをＣrootとして生成する。次に、葉クラスタとして属性ａｍの値がｖｍｖである事例の集合を図示のように生成する。

【００３０】図３の（ｂ）は、初期クラスタの具体例を示す。ここで、全ての事例を含むクラスタＣrootとしてクラスタ“家具”を生成する。次に、葉クラスタとして、属性“脚の数”の値が“０”、“３”を持つものをクラスタ“脚の数＝０”、“脚の数＝３”として図示のように生成する。同様に、葉クラスタとして、属性“材質”の値が“木”を持つものをクラスタ“材質＝木”として図示のように生成する。

【００３１】以上のように全ての事例を含むクラスタＣrootを生成し、葉クラスタとして属性の値毎に事例の集合を生成することによって、事例集合から初期クラスタを生成することが可能となる。

【００３２】図４は、本発明の説明図（その２）を示す。これは、初期クラスタから関連の強い複数のクラスタから１つの新たなクラスタを生成するときの概念を説明する図である。

【００３３】図４の（ａ）は、クラスタ階層中から選択されたペアの２つのクラスタＣi、Ｃjを示す。図４の（ｂ）は、和集合を新クラスタとする方法の例を示す。この２つのクラスタＣi、Ｃjの和集合を新たなクラスタＣｍとする場合には、当該２つのクラスタＣi、Ｃjの和集合の新たなクラスタＣｍを図示のようにペアの上位にリンク付けする。

【００３４】図４の（ｃ）は、積集合を新クラスタとする方法の例を示す。この２つのクラスタＣi、Ｃjの積集合を新たなクラスタＣｍとする場合には、当該２つのクラスタＣi、Ｃjの積集合の新たなクラスタＣｍを図示のようにペアの下位にリンク付けする。

【００３５】図４の（ｄ）は、和集合を模式的に示した図である。和集合（図４の（ｂ）の場合）は、

$|C_i| + |C_j|$

で表現され、クラスタＣiとクラスタＣjとのそれぞれの斜線で示す部分の和となる。

【００３６】図４の（ｅ）は、積集合を模式的に示した図である。積集合（図４の（ｃ）の場合）は、

$|C_i \cap C_j|$

で表現され、クラスタＣiとクラスタＣjとのそれぞれの斜線で示す重なる部分となる。

6

【００３７】図５は、本発明のクラスタ階層例を示す。これは、図３の（ｂ）の初期クラスタについて、既述した和集合の場合の新たなクラスタ　、および既述した積集合の場合の新たなクラスタ　を生成した後のクラスタ階層例である。

【００３８】例えば左側の　の和集合のクラスタは、クラスタ“脚の数＝３”とクラスタ“材質＝木”の２つのペアの和集合として新たなクラスタ　“脚の数＝３　ｏｒ材質＝木”を生成して上位にリンクしたものである。

【００３９】同様に、例えば右側の　の積集合のクラスタは、クラスタ“材質＝プラスティック”とクラスタ“形＝四角”の２つのペアの積集合として新たなクラスタ　“材質＝プラスティック　ａｎｄ　形＝四角”を生成して下位にリンクしたものである。

【００４０】以上のように、クラスタ集合中から関連するペアを見つけ、その和集合／積集合を上位／下位にリンク付けすることにより、関連の大きいクラスタペアから新たなクラスタを生成することが可能となる。

【００４１】図６および図７は、本発明のシステム動作説明フローチャートを示す。これは、既述した図２の詳細なシステム動作を説明するフローチャートであって、図２のＳ１ないしＳ１３に対応する部分を左端に記載する。

【００４２】図６において、Ｓ２１は、事例集合を入力する（Ｓ１）。Ｓ２２は、パラメタｋと、クラスタ制限数Ｌを決定する。Ｓ２３は、ｍ＝０と初期設定する。これは、新たなクラスタＣmを生成するときの変数ｍの値を初期化する。

【００４３】Ｓ２４は、ｉ＝０と初期設定する。Ｓ２５は、ｊ＝０と初期設定する。Ｓ２６は、属性ａｉが値ｖｉｊである全事例からなるクラスタＣｍを作る。

【００４４】Ｓ２７は、ｍ＝ｍ＋１する。Ｓ２８は、ｊ＝（ａｉの取る値の数－１）か判別する。ＹＥＳの場合には、Ｓ３０に進む。一方、ＮＯの場合には、Ｓ２９でｊ＝ｊ＋１し、Ｓ２６に戻り、次のクラスタＣｍを作成することを繰り返す。

【００４５】Ｓ３０は、ｉ＝（全属性数－１）か判別する。ＹＥＳの場合には、Ｓ３２に進む。一方、ＮＯの場合には、Ｓ３１でｉ＝ｉ＋１し、Ｓ２５に戻り繰り返す。Ｓ３２は、Ｐｍａｘ＝０、かつＣｍａｘ＝０と初期設定する。

【００４６】Ｓ３３は、ｉ＝０と初期設定する。Ｓ３４は、ｊ＝ｉ＋１する。Ｓ３５は、クラスタＣiとクラスタＣjに共通して含まれる事例数$|C_i \cap C_j|$を数える。

【００４７】Ｓ３６は、ＣiとＣjが独立な場合の推定共通事例数e_{ij}とその標準偏差d_{ij}を計算する。図７のＳ３７は、$|C_i \cap C_j| > e_{ij} + K \cdot d_{ij}$か判別する。ＹＥＳの場合には、Ｓ３８に進む。ＮＯの場合には、Ｓ４０に進む。

【００４８】Ｓ３８は、$P_{max}<(|C_i \cap C_j|)/(|C_i|+|C_j|)$か判別する。ＹＥＳの場合には、Ｓ３９に進む。ＮＯの場合には、Ｓ４０に進む。Ｓ３９は、$P_{max}=(|C_i \cap C_j|)/(|C_i|+|C_j|)$ $C_{max} \leftarrow (C_i, C_j)$を行う。

【００４９】Ｓ４０は、ｊ＝ｍ－１か判別する。ＹＥＳの場合には、Ｓ４２に進む。ＮＯの場合には、Ｓ４１でｊ＝ｊ＋１し、Ｓ３５に戻り繰り返す。Ｓ４２は、ｉ＝ｍ－２か判別する。ＹＥＳの場合には、Ｓ４４に進む。ＮＯの場合には、Ｓ４３でｉ＝ｉ＋１し、Ｓ３４に戻り繰り返す。

【００５０】Ｓ４４は、$C_{max} \neq 0$か判別する。ＹＥＳの場合には、Ｓ４５に進む。ＮＯの場合には、終了する。Ｓ４５は、C_{max}からＣｍをオペレータにより作る。

【００５１】Ｓ４６は、ｍ＝ｍ＋１する。Ｓ４７は、ｍ＝Ｌか判別する。ＹＥＳの場合には、終了する。ＮＯの場合には、Ｓ３２に戻り繰り返す。

【００５２】

【発明の効果】以上説明したように、本発明によれば、各属性の各値についてその値を持つ事例を集めてクラスタを生成し、クラスタ集合中で共通して出現する事例の数の多いクラスタは関連が高いと見做して複数クラスタから新たなクラスタを生成することを繰り返してクラスタ階層を生成する構成を採用しているため、クラスタ生成時間を短くして１つの事例が複数のクラスに属することを許すクラスタ階層を生成することができる。これらにより、

（１）　属性の値毎にクラスタを１つづつ生成し、共通して出現する事例の数が多いクラスタ同士から新たなクラスタを生成し、１つの事例が複数のクラスに属するよ

うな場合にも対応することが可能となる。

【００５３】（２）　クラスタ階層の生成に必要な処理時間は、階層中のクラスタ数の３乗と全事例数に比例する。通常、複雑すぎるクラス集合は、知識として利用し難いため、必要とされない。そこで、パラメタとして設定される階層中のクラスタ数を、事例数に比べて小さい値にできる。その結果、比較的に短い時間でクラスタの生成が可能となる。

【００５４】（３）　上記（１）および（２）により、１つの事例が複数の概念に属するような大規模な事例集合に対して、効果的にクラスを構成するクラスタを学習することが可能となる。

【図面の簡単な説明】

【図１】本発明のシステム構成図である。

【図２】本発明の動作説明フローチャートである。

【図３】本発明の説明図（その１）である。

【図４】本発明の説明図（その２）である。

【図５】本発明のクラスタ階層例である。

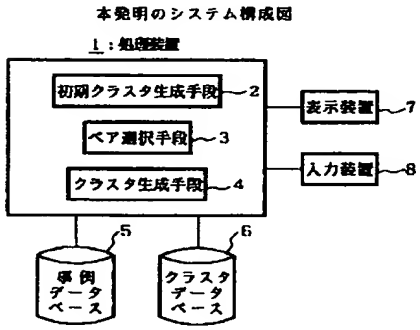
【図６】本発明のシステム動作説明フローチャート（その１）である。

【図７】本発明のシステム動作説明フローチャート（その２）である。

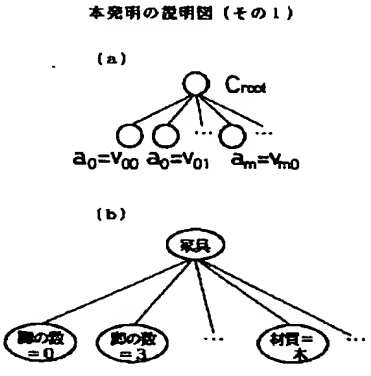
【符号の説明】

１：処理装置
２：初期クラスタ生成手段
３：ペア選択手段
４：クラスタ生成手段
５：事例データベース
６：クラスタデータベース
７：表示装置
８：入力装置

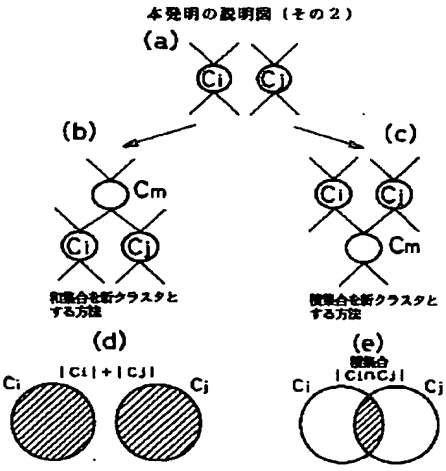
【図１】



【図３】

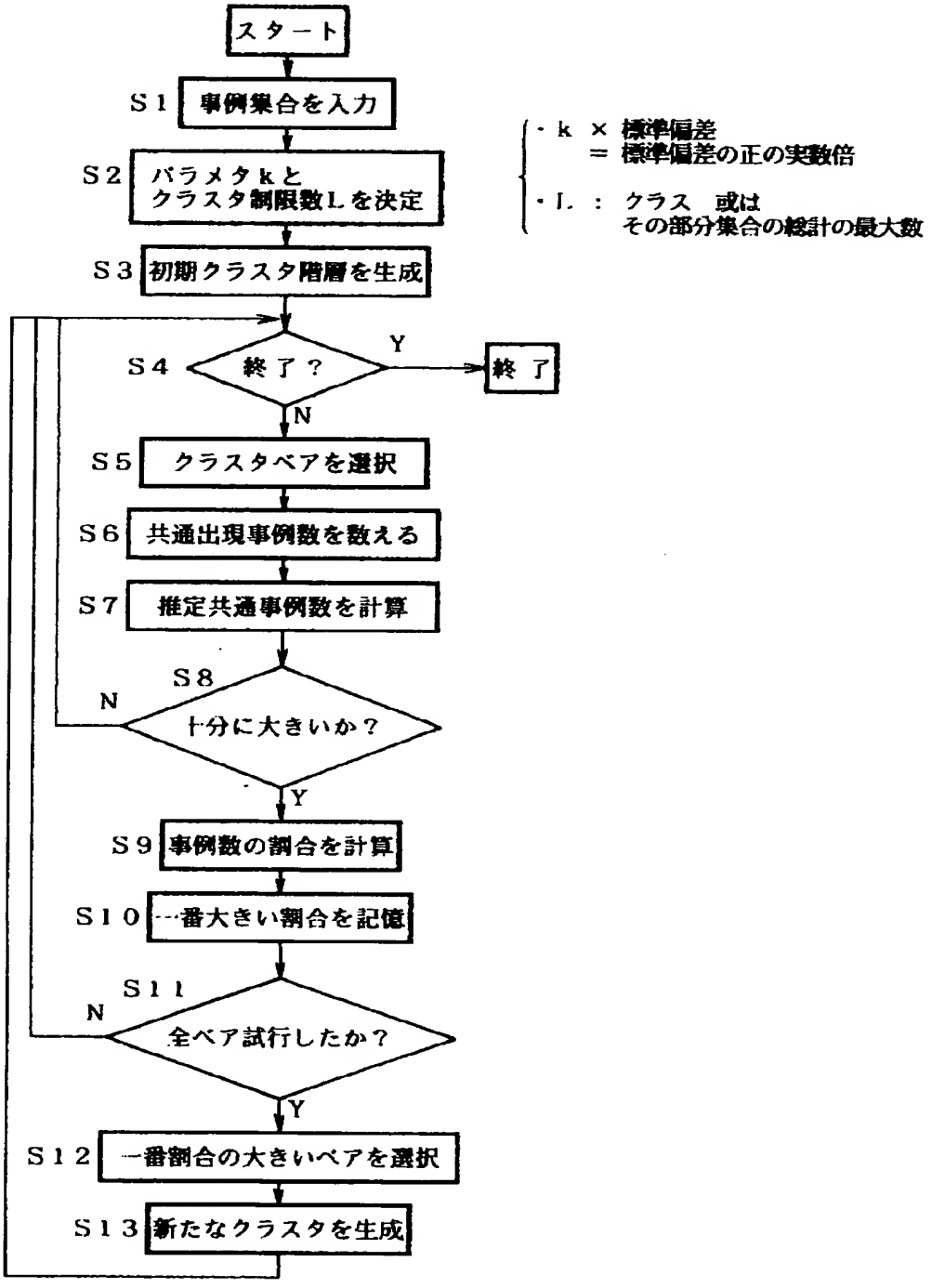


【図４】



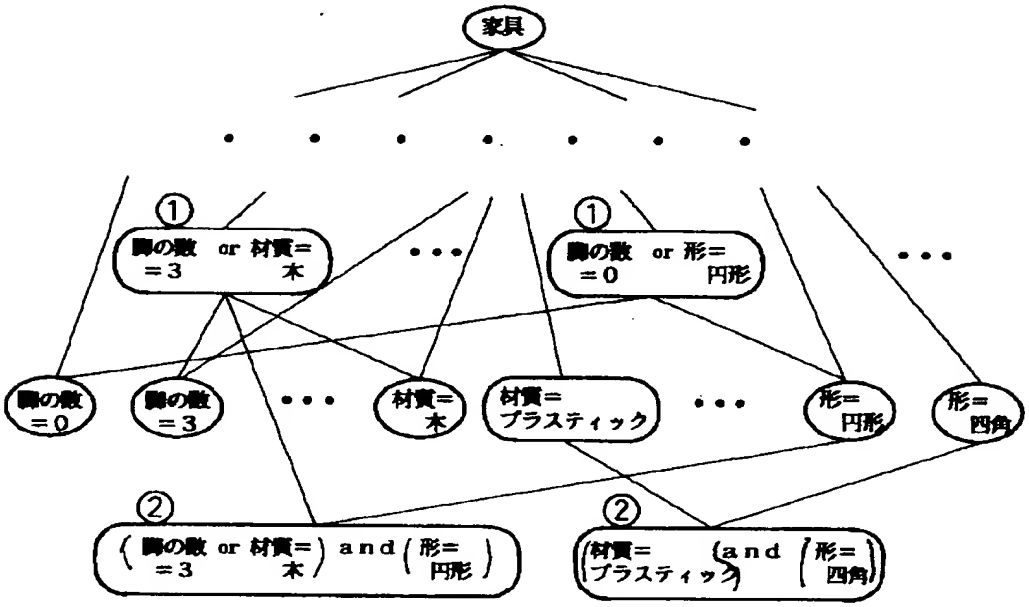
【図２】

本発明の動作説明フローチャート



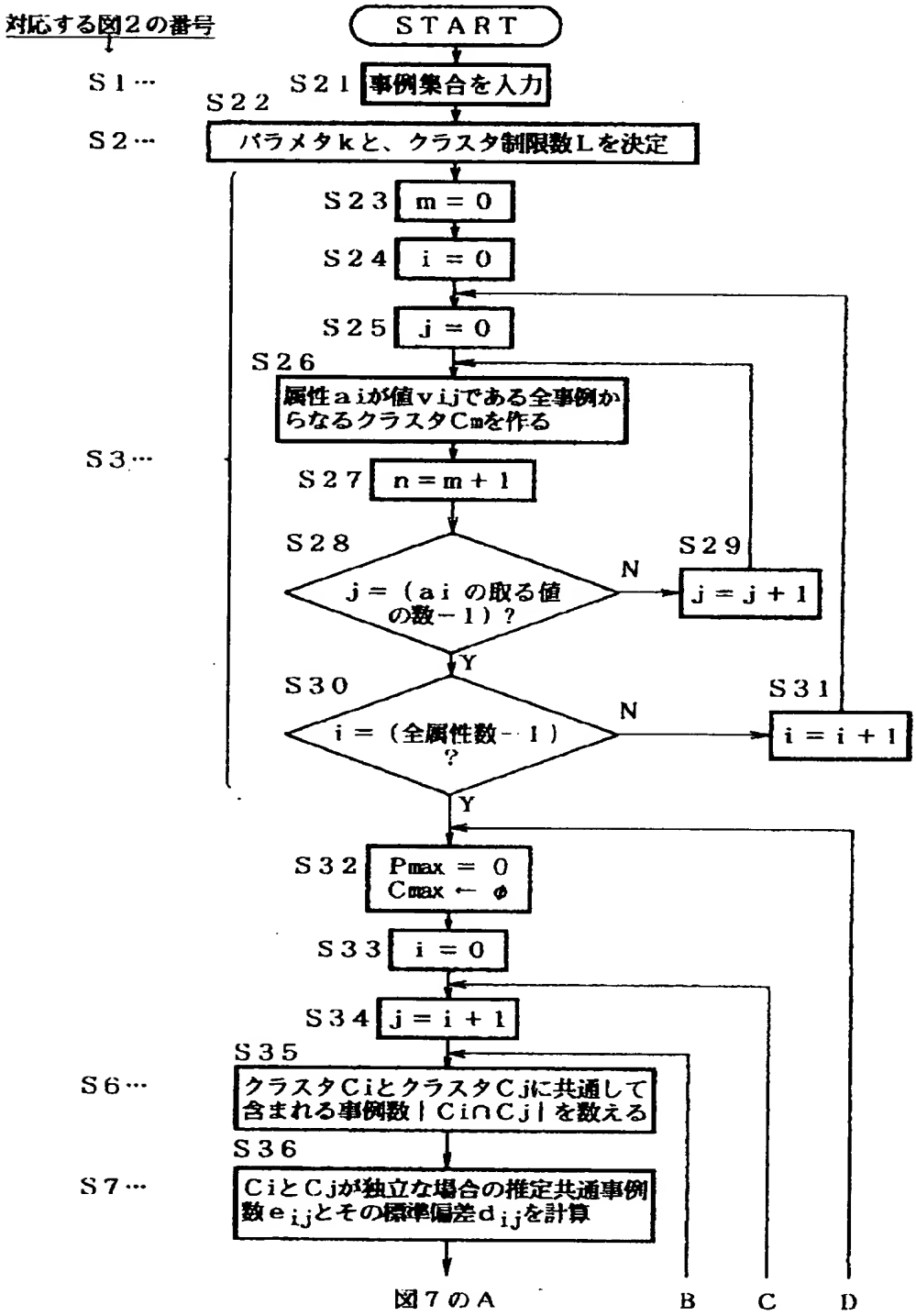
【図５】

本発明のクラスタ階層例



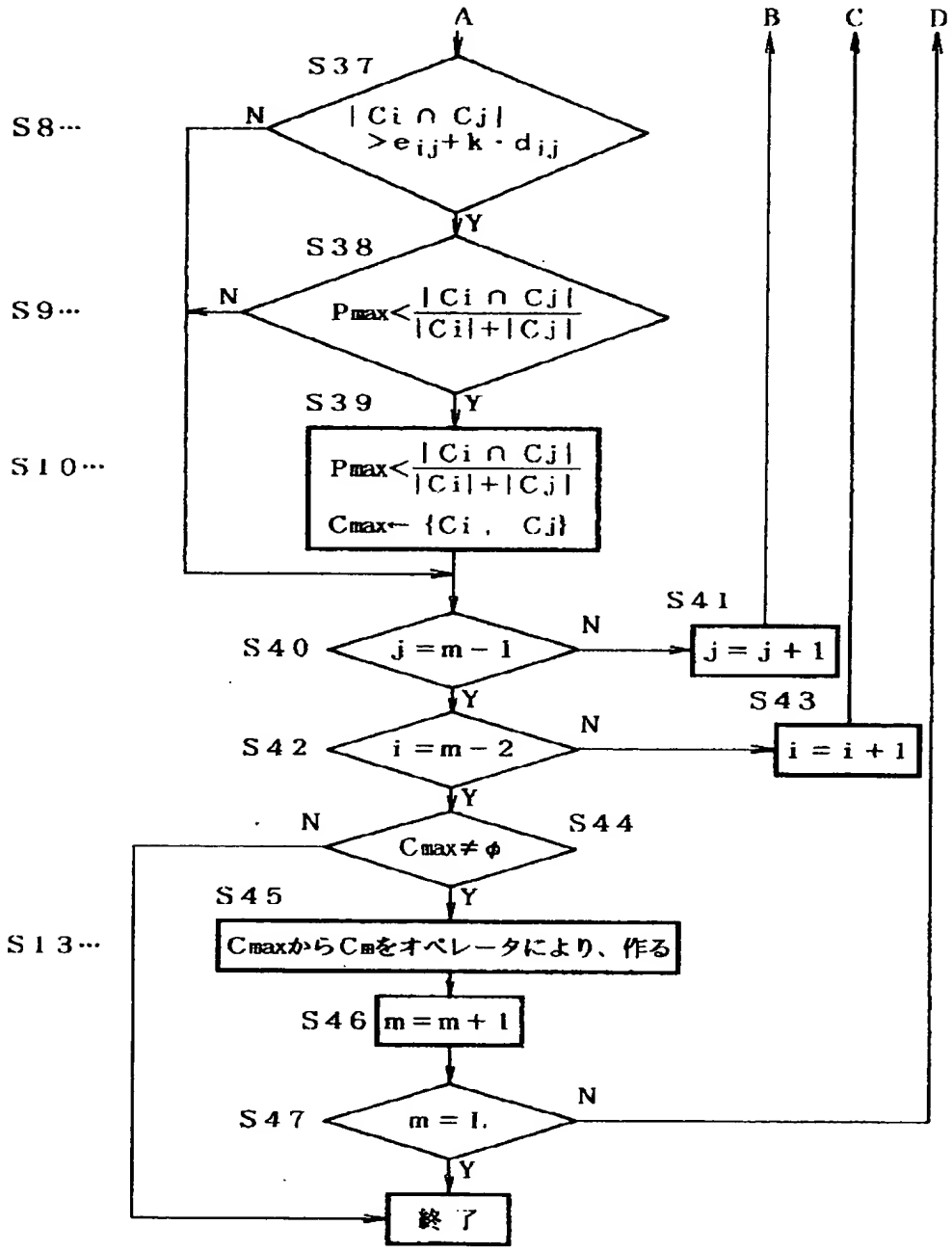
【図６】

本発明のシステム動作説明フローチャート（その１）



【図７】

本発明のシステム動作説明フローチャート（その２）



フロントページの続き

(72)発明者　岡本　青史
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内

(72)発明者　佐藤　理
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内